Using multiple imputations to handle missing data on exposure variables: a national multicenter cohort study of heroin dependents to evaluate the impact of treatments on mortality

Patrizia Schifano[1], Marcello Pagano[2], Alexandros Gryparis[3]

[1] Department of Epidemiology, ASL RME, Via di Santa Costanza 53, 00198 Roma, Italy
[2] Harvard School of Public Health, Harvard University, Boston, Massachusetts, USA
[3] Department of Applied Mathematics, University of Crete, Greece

Corresponding Author:
Patrizia Schifano, Department of Epidemiology, ASL RME
Via di Santa Costanza, 53 - 00198 Rome, Italy
E-mail: schifano@asplazio.it

Summary

Objectives. Missing values represent a common issue in longitudinal studies and different methodologies have been proposed to approach it. We consider the particular case of longitudinal studies where vital status follow-up is more extended in time than exposure measurement follow-up. We use the example of a prospective cohort Italian study to evaluate the impact of treatments for heroin dependence on mortality from overdose, to show how multiple imputation (MI) helps in properly treating the lack of information, offering a better solution than the more common list-wise deletion method.

Methods. The study recruited 10,258 subjects and detailed information on each treatment episode was collected for an 18-month period. Vital status was also updated at 24 months from recruitment. To account for the missing information on treatment for the one-year period between the end of the study and the last vital status ascertainment, we apply a multiple imputation approach based on all the available information included in the observed data.

Results. The number of deaths included in the complete data set is 89% higher than that in the observed data set. The Cox model applied to the complete data set produces higher point estimates, but comparable 95%CI. The two validation models show the robustness of the adopted method.

Conclusions. The imputation model slightly changes the results of the list-wise deletion analysis, but has definitely made these results more valid and reliable. Multiple imputation should be considered more often as an applicable choice in the treatment of missing values in cohort studies where collecting information on exposure data might be very expensive, while the information on outcomes could be easily updated from administrative databases.

KEY WORDS: *multiple imputation, missing values, cohort study, exposure*.

Introduction

Missing values are an issue of major concern in observational studies, especially so in longitudinal studies. Two different types of missing value may affect a longitudinal observational study: baseline and follow-up information. In fact, in longitudinal studies it may be difficult to collect information at baseline, since subjects may refuse to answer the questionnaire entirely, or in part, or it may be difficult to collect complete information from administrative databases. Moreover, in a longitudinal study, subjects may drop-out, be lost to follow-up, or refuse to participate in subsequent waves of data collection (1).

A certain amount of losses at follow-up is always taken into account in such studies; if the process that generates the losses is not completely random, this may cause selection bias, and the analyst should try to correct it. Especially when there are relatively

many missing data, the results of this selection bias may be severe, leading to biased estimators.

The standard approach, which is a default option in many statistical packages, is to restrict the analysis to subjects with no missing values in the specific sets of variables. This method, called available-case analysis, can yield biased estimates, in the presence of selection bias. Furthermore, when multivariate analyses are involved, including descriptive analyses, a popular approach leads to a listwise "deletion", which is defined as excluding all the subjects that have any missing information in any of the variables used in at least one analysis. This approach leads to both biased parameter estimators and a loss of power.

Different methods have been developed to handle missing values. The most popular in the literature are weighted estimation procedures (2, 3), maximum likelihood estimates obtained through specialized numerical methods, such as EM (4, 5), applicable in fully parametric models, and multiple imputations (MI) (6, 7). All these methods have advantages and drawbacks, but all of them possibly lead to unbiased estimation, if used properly. However, MI is often preferred because, although it represents an approximate solution, it has good properties and is readily available in many statistical software packages. MI is a simple, widely applicable and easily available approach that represents a good solution to the problem of missing data. The most important drawback of MI is that is computationally expensive, although that is not of major concern given the latest computing systems. Many examples of MI are present in the biomedical literature, applied either to baseline missing information (8-10) or to follow-up missing information (10-12).

In this paper, we apply an MI approach to a special case of the follow-up missing data problem: a cohort study in which, by design, treatment information has been collected for a shorter time than information about outcomes. Hence, for a time period we have data on outcomes, but not on the given treatments of interest (exposure). We have considered treatment data in this time period as missing data. To handle the missingness, we use an MI model. Our data come from a prospective, national, cohort study to evaluate the impact of treatments for heroin dependence on mortality from overdose, in Italy (13).

Methods and data

Data

The VEdeTTE study recruited 10,454 heroin users at 115 Public Treatment Centres (PTC) of 554 (23%) within the National Health Service in Italy, at the time of the study. Enrolment included incident subjects that started treatment for the first time at a specific PTC during the study period, "re-entered" subjects that were not in treatment at the beginning of the study but treated by the centre in the past, and "prevalent" subjects that had an open treatment at the beginning of the study. Clients were recoded as "old", if they had been in treatment for more than 6 months at the start of the study, and as "new", if they had been in treatment for less than 6 months, or not yet in treatment, at the beginning of the study. The subjects were followed over an 18-month period, between September 1998 and March 2000. Clinical history and personal information were collected at the intake interview, and each treatment episode over the study period was recorded. Patients could have multiple periods of the same treatment or different treatments. Treatment regimens included methadone maintenance, methadone detoxification, admission in a therapeutic community (residential or semi-residential), other pharmacological detoxification treatments (including naltrexone, in-patient detoxification, detoxification with non-opiate drugs, therapy with psychotropic drugs), and psychosocial types of treatment (including psychotherapy, counselling, social advice and job guidance). Periods without treatment was recorded as well, and labelled as "treatment free" (14).

In order to investigate the association between treatment and overdose mortality, we analyzed the data from 10,258 patients (98% of the whole cohort) for which treatment information was available. Person-years at risk were calculated from the start of the first treatment within the study period to the end of the eighteen-month study period or until the date of death, whichever came first. Vital status was assessed at each treatment centre and through the Registry Office of the last municipality of residence and was reported for 96.3% of the subjects. The last date of vital status ascertainment was March 2001 (14).

Missing data problem

In order to investigate the association between overdose mortality and the different treatments, information about treatment updated to the last follow-up ascertainment is needed. In our data, because of the study design, there was a one-year time gap between the last day in which the information about treatment was collected and the day in which vital status was assessed. One-hundred deaths occurred within the 18-month study period (41 due to overdose) and other 90 deaths (29 due to overdose) occurred in the year between the end of the study and the last mortality follow-up update. This almost doubling of events could not be ignored.

Initially, we excluded from the analysis the 90 cases who died after the end of the study period (14). Since time of death may well be related to the severity of the addiction, this approach introduces selection bias and, moreover, results in a loss of statistical power. In our case, the missing data were generated by design, since the follow-up of the subjects ended at the end of the eighteen month period. In the notation of Little and Rubin (7), our data are missing completely at random (MCAR). Hence, it does not seem plausible that we introduce selection bias, by doing an available-case analysis. Nevertheless, in this specific case, there is still a risk of selection bias. Indeed, as a consequence of the missing information on treatments, we would have had to exclude from the analysis those who died later. These subjects might differ systematically from those included in the analysis. Furthermore, we reduce the sample size (with a particularly strong impact on the number of outcomes included in the analysis), and therefore lose statistical power. To take into account both these problems, we decided to handle the missing information through MI (6). This approach properly uses the partial information available on treatments and imputes the exposure variable for the one-year period in which it was not collected. Such an approach assumes that the missing data were missing at random (MAR) (7).

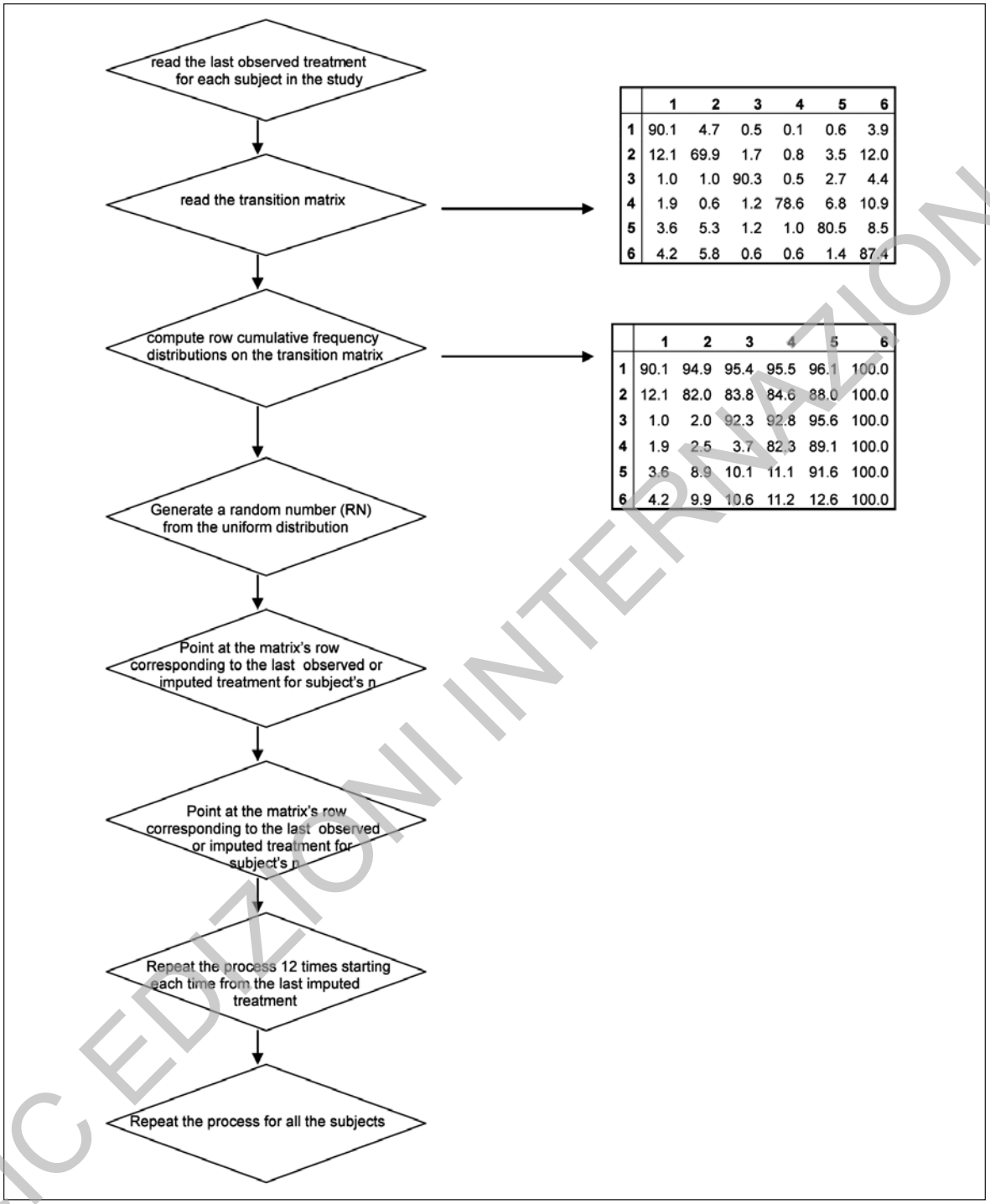
The MI model

First, we assume that the process of switching from one treatment to another behaves as a Markovian chain with a one month cycle. Therefore, the probability of switching to a certain treatment only depends on the current treatment, and given the last treatment all other past treatment information is independent of the new treatment.

In order to impute the missing data, we estimate a probability transition matrix, from the observed treatment distribution. All treatments observed along the 18 months of the study were split in one-month long treatment sections and, for the months that included more than one treatment episode, only the one of longest duration was considered. Using these fractions of treatments we estimated the overall probabilities of transitioning from treatment regimen at time "i" to treatment regimen at time "i+1", in a one-month period of time.

The transition matrix was estimated for old and new patients separately, since the probability of switching treatments was known to be different between the two groups of subjects. A different transition matrix was produced for each couple of subsequent months. The 18 estimated matrices proved to be very similar; hence we used the overall transition probability matrix that was calculated for the whole 18-month period. All analyses were performed in R, and for the MI approach we programmed our own code (Figure 1). The patients who did not die within the first 18-month of the study period were assigned to 12, at most, additional treatments of one-month length. These additional treatments are meant to cover the time period between the last observed treatment and the last vital status ascertainment. The data imputation algorithm results in a "complete" dataset. We then fit a Cox model, to estimate the hazard ratios of surviving when exposed to each one of the possible treatments compared to being out of treatment. The above process was repeated 600 times, and therefore 600 completed datasets were created. The number of repetitions was chosen in order to achieve the desired level of accuracy.

Along with the indicator variables for the treatment effect, our aim is to account for other important covariates that could possibly confound or affect the association of interest. Only baseline covariates are available. To select among those, we used available scientific knowledge from previous studies, along with a model selection procedure. For the latter, we fit the model on the whole population, without adjusting for treatment regimens. Hence, we had to fit



	1	2	3	4	5	6
1	90.1	4.7	0.5	0.1	0.6	3.9
2	12.1	69.9	1.7	0.8	3.5	12.0
3	1.0	1.0	90.3	0.5	2.7	4.4
4	1.9	0.6	1.2	78.6	6.8	10.9
5	3.6	5.3	1.2	1.0	80.5	8.5
6	4.2	5.8	0.6	0.6	1.4	87.4

	1	2	3	4	5	6
1	90.1	94.9	95.4	95.5	96.1	100.0
2	12.1	82.0	83.8	84.6	88.0	100.0
3	1.0	2.0	92.3	92.8	95.6	100.0
4	1.9	2.5	3.7	82.3	89.1	100.0
5	3.6	8.9	10.1	11.1	91.6	100.0
6	4.2	9.9	10.6	11.2	12.6	100.0

Figure 1. Description of the algorithm used to create the imputes.

the model just once, on the original data. Covariates that proved to be statistically significant at a level of 0.10 were included. All treatments, age and gender were included in the model, independently of their statistical significance. The selection model procedure is exactly that followed for the available-data analysis. To summarize the coefficients of interest, we pooled all the parameter estimates using the mean of the estimates obtained in each imputation. Their standard errors are computed by taking the square root of the sum of the between imputations and within imputation variances (15).

Model validation

To validate our imputation model, we use two different strategies. First, we implement a more complex probabilistic model: we assume that for the last one-year period of the study, the probability of getting a specific treatment regimen depends on the treatments one receives in the last two months of the observation period, rather on the last month only. We refer to this model as the second probabilistic model. Second, we applied both the first and the second probabilistic model to the first nine months of the observed study period, in order to predict the last nine months of the observed period. We then compared the observed and the predicted treatments. In this case, the transition matrix is estimated using only the first nine months of the observed period.

Results

We included 10,258 subjects in the analysis. The imputation process allows us to use all observed deaths (189, 70 from overdose), rather than just those that correspond to the first 18-month period (100, 41 from overdose). The 29 overdose cases that correspond to the last year, were younger (mean age at enrollment 30.8 (SD 5.7) vs 34.9 (SD 5.5)), with a lower prevalence of HIV sero-positives (3.4 % vs 26.8 % of HIV-positive), and in treatment for a longer period at the beginning of the study (34.5% vs 24.4% patients in treatment for more than 6 months).

By using MI, we imputed 10,116 person years. Table 2 shows the distributions of imputed and observed person years. The most important differences are the higher percentage of person years spent out of treatment and the decrease in the percentage of person years spent in methadone detoxification in the complete data set when compared to the observed one.

Comparing the results of the Cox survival analysis on the completed and observed data set, we observe an increase in the precision of all the baseline covariates (that were fixed throughout the imputation process) estimates. Moreover, HIV serological status is no longer statistical significant, probably due to

Table 1. Socio-demographic and drug use characteristics of the study population, VEdeTTE study.

	All population		Deaths before the end of the study		Deaths after the end of the study		All deaths		Deaths for overdose before the end of the study		Deaths for overdose after the end of the study	
	N. (10258)	%	N. (100)	%	N. (89)	%	N. (189)	%	N. (41)	%	N. (29)	%
Social characteristics												
Gender												
men	8779	85.6	80	80.0	72	80.9	152	0.8	31	75.6	25	86.2
women	1479	14.4	20	20.0	17	19.1	37	0.2	10	24.4	4	13.8
Living Status												
alone	947	9.2	19	19.0	26	29.2	45	23.8	9	22.0	8	27.6
with parents	5617	54.8	57	57.0	37	41.6	94	49.7	25	61.0	12	41.4
with partner and/or with child	2756	26.9	19	19.0	17	19.1	36	19.0	4	9.8	5	17.2
other	890	8.7	5	5.0	8	9.0	13	6.9	3	7.3	4	13.8
missing	*48*	*0.5*	*0*	*0.0*	*1*	*1.1*	*1*	*0.5*	*0*	*0.0*	*0*	*0.0*
Education												
never went to school	93	0.9	2	2.2	1	1.1	3	1.6	1	2.4	0	0.0
five years of education	1650	16.1	16	18.0	13	14.6	29	15.3	4	9.8	5	17.2
eight years of education	6071	59.2	57	64.0	60	67.4	117	61.9	23	56.1	20	69.0
secondary school	2317	22.6	22	24.7	14	15.7	36	19.0	12	29.3	4	13.8
higher education	81	0.8	2	2.2	1	1.1	3	1.6	1	2.4	0	0.0
missing	*46*	*0.4*	*1*	*1.1*	*0*	*0.0*	1	0.5	*0*	0.0	*0*	0.
Age at enrolment												
average (SD)	31.5 (6.2)		35.4 (6.1)		33.1 (6.5)		34.3 (6.4)		34.9 (5.5)		30.8 (5.7)	

Table 1. Socio-demographic and drug use characteristics of the study population, VEdeTTE study.

	All population		Deaths before the end of the study		Deaths after the end of the study		All deaths		Deaths for overdose before the end of the study		Deaths for overdose after the end of the study	
	N. (10258)	%	N. (100)	%	N. (89)	%	N. (189)	%	N. (41)	%	N. (29)	%
Drug addiction history												
Injectors												
yes	7431	72.4	88	88.0	74	83.1	162	0.9	34	82.9	21	82.8
no	1819	17.7	7	7.0	6	6.7	13	0.1	3	7.3	2	6.9
missing	*1008*	*9.8*	*5*	*5.0*	*9*	*10.1*	*14*	*0.1*	*4*	*9.8*	*3*	*10.3*
Imprisonment												
yes	1472	14.3	12	12.0	12	13.5	24	0.1	5	12.2	2	6.9
no	8178	79.7	81	81.0	72	80.9	153	0.8	36	87.8	26	89.7
missing	*608*	*5.9*	*7*	*7.0*	*5*	*5.6*	*12*	*0.1*	*0*	*0.0*	*1*	*3.4*
Overdose in the past												
yes	4170	40.7	56	56.0	53	59.6	109	0.6	26	63.4	17	58.6
no	5967	58.2	44	44.0	34	38.2	78	0.4	15	36.6	11	37.9
missing	*121*	*1.2*	*0*	*0.0*	*2*	*22*	*2*	*0.0*	*0*	*0.0*	*1*	*3.4*
Co-abuse of cocaine												
yes			35	35.0	35	39.3	70	0.4	12	29.3	9	31.0
no			60	60.0	48	53.9	108	0.6	28	68.3	19	65.5
missing			*5*	*5.0*	*6*	*6.7*	*11*	*0.1*	*1*	*2.4*	*1*	*3.4*
Age (yrs) at first injecting heroin use average (SD)	20.8 (4.7) n=9074		20.2 (4.4) n=95		19.93 (4.5) n=84				20.8 (4.3) n=39		20.2 (4.3) n=27	
Clinical features												
Type of client (a)												
in the treatment ≥ 6 months at the begin of the study	2612	25.5	31	31.0	22	24.7	53	0.3	10	24.4	10	34.5
in the treatment < 6 months at the begin of the study	7646	74.5	69	69.0	67	75.3	136	0.7	31	75.6	19	65.5
Type of client (b)												
new at the treatment centre	1215	11.8	2	2.0	9	10.1	11	5.8	2	4.9	3	10.3
already known at the treatment centre	9043	88.2	98	98.0	80	89.9	178	94.2	39	95.1	26	89.7
Psychiatric diagnosis in the past												
yes	1380	13.5	30	30.0	20	22.5	50	0.3	13	31.7	10	34.5
no	6746	65.8	50	50.0	47	52.8	97	0.5	20	48.8	15	51.7
missing	*2132*	*20.8*	*20*	*20.0*	*22*	*24.7*	*42*	*0.2*	*9*	*22.0*	*4*	*13.8*
HIV												
yes	841	8.2	43	43.0	26	29.2	69	0.4	11	26.8	1	3.4
no	5966	58.2	33	33.0	41	46.1	74	0.4	19	46.3	21	72.4
missing	*3451*	*33.6*	*24*	*24.0*	*22*	*24.7*	*46*	*0.2*	*11*	*26.8*	*7*	*24.1*

* In the last 12 months.

the very different distribution of added deaths relative to this covariate.

Furthermore, the "methadone maintenance " and the "psycho-social" treatment estimates had an increased variance (the 95% CI is almost double) , due to the imputation variability component, while the variance components of the rest of the treatment regimens remained more or less the same as before. The point estimates are generally higher for all treatments, but the CIs overlap a lot, indicating that the results of the completed dataset and the available case are quite similar. It is important that with the MI model we are able to estimate the therapeutic community effect, while there were no deaths attributed to this treat-

Table 2. Variables associated with overdose death among heroin users. Results from the observed and complete data sets, VEdeTTE study.

	Deaths				Person-years				RR*					
	Complete Dataset N=70	Observed Dataset N=41	Imputed	%	Observed	%	Total	%	Complete Dataset	RR	95% IC	Observed Dataset	RR	95% IC
Treatment														
Out of treatment	–	31	4067	40	2914	22	6980	30	1.00			1.00		
Methadone maintenace	–	7	3602	36	5751	44	9353	40	0.21	0.11	0.39	0.10	0.04	0.24
Methadone detoxification	–	1	525	5	1496	11	2020	9	0.13	0.03	0.56	0.07	0.01	0.50
Other pharmacological	–	1	257	3	423	3	679	3	0.84	0.26	2.73	0.35	0.05	2.58
Psychosocial	–	1	1011	10	1349	10	2360	10	0.33	0.13	0.84	0.08	0.03	0.32
Therapeutic Comunity	–	0	655	6	1189	9	1844	8	0.23	0.07	0.74	–	–	
Total			10116	100	13121	100	23238	100						
Gender														
Male	56	31	8659.5		11606		20265.4		1.00			1.00		
Female	14	10	1456.7		1932		3388.7		1.48	0.82	2.68	2.03	0.98	4.18
Age														
Five years increasing									1.04	1.00	1.08	1.36	10.7	1.73
Psychiatric comorbidity														
No	23	13	6656.4		7982		14638.24		1.00					
Yes	35	20	1365.7		1861		3226.48		2.96	1.74	5.05	2.76	1.36	5.62
HIV serological status														
Negative	12	11	5887.6		7951		13838.1		1.00			1.0		
Positive	40	19	829.5		1156		1985.6		1.71	0.87	3.35	2.88	1.32	6.24
Non-fatal overdose														
No			5887.6		7782		13669.4		1.00			1.00		
Yes			4117.3		5603		9719.9		2.05	1.25	3.35	2.09	1.10	3.97

ment in the observed data. This treatment regimen shows a significant protective effect, that is consonant with what is otherwise known about this treatment.

Model validation

Using the last two observed treatments to estimate the transition matrix produced results that are consistent with those of the chosen imputation model (Table 3). The second procedure, instead, shows that the adopted imputation model predicts sufficiently accurately the person years for "methadone maintenance", "other pharmacological" and "therapeutic community", while it overestimates the time spent in "methadone detoxification" and it underestimates the time spent "out of treatment" (Table 3).

Discussion

Summary of results

Imputation allows us to incorporate all the events that occurred in the cohort at the latest time of observation, increasing the number of overdose deaths analyzed from 41 to 70. Overall person years increased from 13,121 to 23,238.

Source of missingness

We apply a well known method for dealing with missing values to a special case of missingness, in which missing values are totally attributable to the study design.

Table 3. Results of the validation models.

					Imputed values for the not observed 12 months				Imputed values for the observed 9 months			
	Original data (all 18 months)		Observed data (second 9 months)		Adopted model		First validation model (last 2 treatments)		Adopted model		First validation model (last 2 treatments)	
	Person years	%	Person years	%	Person years	%	Person years	%	Person years	%	Person years	%
Treatment												
Out of treatment	2914	22	1436	35	4067	40	4115	41	1272	20	1600	25
Methadone maintenace	5751	44	186	29	3602	36	3594	36	1594	25	1543	25
Methadone detoxification	1496	11	694	17	525	5	577	6	1722	27	1531	24
Other pharmacological	423	3	125	3	257	3	231	2	269	4	239	4
Psychosocial	1349	10	383	9	1011	10	955	9	1017	16	958	15
Therapeutic Comunity	1189	9	242	6	655	6	643	6	402	6	406	6
Total	13121	100	4066	100	10116	100	10116	100	6277	100	6277	100

It is not uncommon in cohort studies that the study design imposes an end to the follow-up period, often due to budget constraints. However there are outcomes that can be followed-up very easily and substantially costless; as is the case of mortality, whenever a mortality register exists.

So, even if the study is not able to follow-up fully the cohort for exposure, it might nevertheless be able to update vital status ascertainment at future points in time, through record linkage procedures with existing databases. Exposure data for this amount of time may be considered missing values. They represent a special case of missing values, because they are generated as a consequence of the study design. Then multiple imputation may represents an important method to "update" exposure data, allowing a more powerful and valid analysis.

Why multiple imputation

We chose multiple imputation among the possible methods to deal with missing values because it has known, good statistical properties (16, 17). One should mention that if the imputation model is seriously flawed in terms of capturing the missing data mechanism, then so will be any analysis based on such imputations. Consequently it must be considered as part of the MI approach to control the validity of the adopted model, by making the best use of what is known about the data mechanism and of the observed data, and by performing an adequate sensitivity analysis (18). In our study, because missing values were generated because of the interruption of the follow-up period, we could not apply any model driven approach to the understanding of the missing values mechanism. However the only possible hypothesis is that these missing values are of the MAR type. Furthermore sensitivity analysis results validate this hypothesis. Some authors however suggest adopting a more complex approach, like a joint-modelling approach, when the necessary assumptions are valid (19).

Nature of missing values and the problem of potential selection bias

It is well known that multiple imputation requires at least MAR missing values (20). In this case we have MCAR missing values, because they are due to the ending of the study, and involved all patients alive at that time. As a consequence we should not be concerned about potential bias. But in this specific case we actually are. Lack of information on treatments after the end of the study means the reduction of the analysis to those who died in the first 18 months; excluding those who died later in time. But those who died in the first 18 months of the study were different from those who died later in more than one respect: they were older, in treatment for a shorter time, and had a higher prevalence of HIV sero-posi-

tivity. So it is important that the group of those dead in the last year of the observational period be included in the analysis, in order to have more representative results.

Problem of power

The deletion of cases method would have required to discarding 29 cases of overdose deaths out of 70 observed at the last vital ascertainment. This implies a loss of about 40% of the events, which makes it very difficult to estimate the treatments effect, and indeed makes it impossible to estimate the effect of therapeutic community because of the absence of cases within that treatment group.

Comparison of estimates precision

It is interesting to note that although the MI method has an extra component estimated in the standard error (the component estimating uncertainty due to imputation of missing data) the standard error from MI is still almost equal to that produced by listwise deletion for most of the treatment effects. This might suggest that the increase of variability due to uncertainty attributable to imputation is compensated for by the increase in power. Furthermore, as expected, we gain precision in the estimate of all the fixed covariates. We also observe a change in the significance of the HIV status covariate estimate; this may be interpreted as a better estimate for this covariate, due to the greater representativeness of cases included in the analysis.

The imputation model

It is well known that for a good MI one should choose a good and inclusive model (18). Our model is based on a strong assumption that the current treatment only depends on the last received treatment. This may be criticized, because often the whole treatment history may be important in determining a patient's future treatments. Nevertheless, it is also true that it is the last received treatment that best describes the situation of the person at the moment, and so it can be used to predict the next treatment.

Also, a more comprehensive model could have been tried. But the weak evidences of an association between treatments distribution and available covariates make it unnecessary.

Furthermore, both the validation models we tried give good and reliable results and reinforce our assumptions that the adopted model adequately describes the missing values mechanism. However, the observed underestimation of the out of treatment PY and the overestimation of the methadone detoxification PY in the second validation approach might be due to the adopted imputation model, in which the joint distribution of observed treatments and their length in time determine the probability to be imputed.

In this specific case we have not been able to use the available procedures present in statistical software packages, because of the nature of our missing values. We had to develop our own code using R. The principle disadvantage of this method is computational inefficiency; the R code took a very long time to run even when we used a high speed processor.

Conclusions

Although the use of more sophisticated and "scientific" methods to deal with missing values are becoming more common, listwise deletion is still a frequently used solution in many epidemiologic studies. Multiple imputation has the great advantage of allowing us to handle separately the incomplete-data problems and the substantive analysis. Furthermore it may be considered relatively friendly and easily applicable.

Our example shows the application of this method made it possible to increase the validity of an analysis that otherwise would have been penalized by both loss of power and potential selection bias. Furthermore, our study represents an example of how MI can be used to impute exposure information in cohort studies; if the validity of this use of MI is corroborated by other examples, it may be considered as a useful tool to update long lasting longitudinal studies, where it is easy to collect data on outcomes, but not on exposure.

Even if multiple imputation requires some more efforts in terms of computational skills and time than listwise deletion, we believe that the results make it

worthwhile, and that it should be used more, especially in the field of epidemiology.

Acknowledgments

Authors wish to acknowledge Dr. Valeria Belleudi for her help in the data analysis and the Vedette Study Group that has made this work possible. Research for this paper was supported in part (MP) by NIH grant EB006195.

References

1. Raghunatan TE. What do we do with missing data? Some options for Analysis of incomplete data. Annu Rev Public Health 2004; 25: 99-117.
2. Little RJA. Survey nonresponse adjustments. International Statistical Review 1986; 4: 139-157.
3. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 1994; 89: 846-866.
4. Becker NG. Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. Stat Methods Med Res 1997; 6 (1): 24-37.
5. Meng XL. The EM algorithm and medical studies: a historical link. Stat Methods Med Res 1997 Mar; 6 (1): 3-23. Review.
6. Little RJA, Rubin D. Statistical analysis with missing data. New York: John Wiley; 1987.
7. Rubin D. Multiple Imputations for nonresponse surveys. New York: Wiley; 1987.
8. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. Stat Med 2001 May 15-30; 20 (9-10): 1541-1549.
9. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. Am J Epidemiol 2003 Jan 1; 157 (1): 74-84.
10. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10: 585-598.
11. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol 2004 Jul 1; 160 (1): 34-45.
12. Brancato G, Pezzotti P, Rapiti E, Perucci CA, Abeni D, Babbalacchio A, Rezza G. Multiple imputation method for estimating incidence of HIV infection. The Multicenter Prospective HIV Study. Int J Epidemiol 1997 Oct; 26 (5): 1107-1114.
13. Bargagli AM, Faggiano F, Amato L, Salamina G, Davoli M, Mathis F, Cuomo L, Schifano P, Burroni P, Perucci CA for the VEdeTTE Study Group. VedeTTE, a longitudinal study on effectiveness of treatments for heroin addiction in Italy: Study protocol and characteristics of study population. Substance Use and Misuse 2006; 41: 1861-1879.
14. Davoli M, Bargagli AM, Perucci CA, Schifano P, Belleudi V, Hickman M, Salamina G, Diecidue R,Vigna-Taglianti F, Faggiano F for the Vedette Group. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE Study, a national multisite prospective cohort study. Addiction 2007; 102 (12): 1954-1959.
15. Schafer JL. Multiple imputation: a primer. Stat. Methods in Medical Research 1999; 8: 3-15.
16. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychological Methods 2001; 6 (4): 317-329.
17. Touloumi G, Babiker AG, Kenward MG, Pocock SJ, Darbyshire JH. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. Stat Med 2003 Oct 30; 22 (20): 3161-3175.
18. Liu M, Taylor JM, Belin TR. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. Biometrics 2000 Dec; 56 (4): 1157-1163.
19. Barnard J, Meng XL. Application of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res 1999; 8: 17-36.
20. Fraser G., Yan R. Guided multiple imputation of missing data. Using a subsample to strengthen the missing-at-random assumptions. Epidemiology 2007; 18: 246-252.